Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 00 (2013) 000–000

Procedia
Social and Behavioral Sciences

www.elsevier.com/locate/procedia

Bamana Reference Corpus (BRC)

Valentin Vydrin

1. Introduction

Corpus building for Sub-Saharan African languages started with a considerable delay. A tangible progress has been attained for Swahili, where a 12,500,000-word corpus was created in Helsinki by Arvi Hurskainen (see, in particular, Hurskainen 2004). There are considerable corpus-building activities for languages spoken in the Republic of South Africa (de Shryver 2002) and for Amharic (Gambäk, Björn 2012; Asker et al. 2012). A Yoruba corpus has been announced on the site of the Osaka University, and a 500,000-word diachronic Yoruba corpus of newspaper texts has been created at LLACAN (Aubry 2010). A 7-million word Hausa corpus has been built at Potsdam University, and information on some other corpora appears sporadically on the Internet. So far, there seems to be no freely accessible African language corpus, apart from the Bambara Reference Corpus (the Helsinki Corpus of Swahili is "semi-open": an external user's access is difficult to establish because of a rather complicated procedure of obtaining a password, although it is not impossible.).

Bambara (Bamana, Bamanankan, ISO-369 Bam) is the most widely spoken language of the Manding language group (Western Mande < Mande < Niger-Congo). It is spoken mainly in Mali (and among the very considerable Malian diaspora) by 12 to 15 million people; of these, about 4 millon are L1 speakers. It has no status of official language, however it is the major language (besides French) on Malian radio and TV, there are periodicals in Bambara, it is broadly used in literacy programs and in primary schools; it is taught at several universities in Europe and the US. In France, Bambara is included into the list of optional languages for the final exams of baccalauréat (school leaving certificate).

For the first time, the idea of the Bambara corpus building was advanced in 2008 (Vydrin 2008). A working group was created in 2009 in St. Petersburg; it included initially Valentin Vydrin (the coordinator of the group), Kirill Maslinsky (a computational linguist), Artem Davydov and Anna Erman (specialists in Bambara). Later on the group was joined by

Selection and peer-review under responsibility of CILC2013.

colleagues from other countries. In April 2012, the "Corpus Bambara de Référence" (Bambara Reference Corpus, or BRC) was made accessible on the Internet.

In the present paper, modalities of the Bambara corpus-building process are discussed. First of all, certain features of the Bambara language structure relevant for corpus-building will be mentioned. Secondly, the set of electronic tools developed for the BRC will be surveyed. Thirdly, the process of the corpus-building work will be described. Finally, characteristics of the BRC at the current stage and some short- and mid-term perspectives of the project will be addressed.

2. Modalities of the Bambara Reference Corpus

2.1. Certain features of the Bambara language structure relevant for corpus-building

- Bambara is a newly-written language; its first official orthography was adapted in 1967, reformed in 1982 (in reality, implementation of the new orthography began in 1987-1990). Spelling rules are very insufficiently specified in official documents, and the written practice of Bambara is characterized by considerable variability, especially concerning word segmentation, vowel length, nasality, and certain other points.
- Bambara is very rarely used on the Internet. The Bambara Wikipedia counts a couple of hundred entries, most of them rudimentary and often written without any respect for the rules of orthography. So far, websites and blogs in Bambara are virtually non-existent.
- Bambara is a tonal language. It has two level tones and a down-drift, the segmental domain of a toneme is a word (in reality, the situation is more complex). There are grammatical tonal rules: (1) a "floating low tone" can be postulated (which manifests itself on the surface as a downstep, i.e., a lower realization of the tone of the subsequent word), it functions as a referential article, and (2) "tonal compactness" (a prosodic incorporation which manifests itself as a spreading of a toneme of a word to the right; this tonal spreading marks certain types of syntactic relation). Tones are never marked in Bambara press and books published in Mali; tonal notation is present in publications of texts by linguists, however, even in the latter case it desperately lacks uniformity.
- Bambara is an isolating language with certain elements of agglutination. Inflectional morphology is scanty, grammatical meanings are mainly expressed by word order and auxiliaries. In the meantime, there is some derivative morphology.
- There is a highly productive POS conversion: verb → noun, adjective → noun, noun → pre-verbal adverb, etc.
- Word composition is an open-ended process. In fact, compounds are formed according to a couple of productive models, and it is impossible to list all such compounds in dictionaries. In reality, only lexicalized tonally compact sequences of words are regarded as "true compounds", despite the fact that non-lexicalized ones are often written (according to the official orthography rules) without spaces.

The enumerated peculiarities of Bambara outline the major difficulties arising before a corpus-builder: insufficient normalization of the written language; paucity of texts in the electronic format; lack of tonal marking (although quite relevant for the grammar and vocabulary) in written texts; scantiness of the inflectional morphology which would facilitate an

automatic analysis of a text; productive word compounding which blurs the segmentation of a text into words. Ambiguity in a Manding text is much more frequent (about 70%) than in an average text in most European languages; see Fig. 1.

2.2. Tools for the Bambara Reference Corpus

In line with the arguments presented in (Sharoff & Nivre 2011), software development efforts of the Bambara corpus building team were primarily directed at the creation of tools for human annotators. A software suite for the corpus operator "Daba" has been developed by Kirill Maslinsky (Maslinsky 2012) for semi-automatic annotation. A Bambara text sample of 102 000 words compiled by Gérard Dumestre was used for testing the software during the initial period.

The need for specific software stems from the idea that the Bambara corpus should be annotated not at the word-level but at the morpheme-level, in a way similar to the interlinear glossing widely used in the study of Mande languages. Available programs for automatic annotation have proved not to be ready for adaptation to morpheme-level annotation of a Bambara text, since most of them rely heavily on the concept of orthographic word and are biased for certain types of morphological structure (namely, inflectional morphology typical of the majority of European languages).

The Daba morphological parser is a language-independent framework dictionary and a rule-based morphological analyzer. Language-specific data used by the parser consists of a dictionary and a list of rules for splitting words into morphemes. Morphotactic rules describe formal orthographic features for extracting morphemes (using regular expressions) and context constraints on combinations of morphemes. Such rules are similar to word formulas used in the SIL Toolbox or context rules used in the constraint grammar approach (VISL CG-3 software).

Daba includes the following components:

- a non-standard Bambara orthography converter. The old Bambara orthography (before 1987) cannot be transformed into the new one by automatic replacements; this problem can be solved by addressing the dictionary. The converter is easily adaptable to other non-standard Bambara orthographies used in some publications;
- a dictionary-based morphological analyzer taking into account combinatorial constraints on inflectional and derivational affixes and stem composition, which helps decrease homonymy in the texts;
- a graphical user interface (GUI) for semi-automatic disambiguation of automatically annotated texts (see Fig. 1);
- a GUI for the introduction of descriptive metadata, designed to classify each text according to a hierarchy of attributes, which are largely based on the EAGLES Recommendations on Text Typology compiled by Sinclair & Ball (1996), cf. (Davydov 2010).

The Daba morphological parser is written in the Python programming language and is distributed under the terms of GNU GPL License. The choice of Python is substantiated by the fact that this language is widely used in the natural language processing community, which facilitates access to this software for the community of computational linguists and its use in other similar projects. Another argument in favour of Python is the availability of considerable libraries of utilities for NLP tasks (NLTK).

In 2010–2012, the following resources were developed:

- a list of standard glosses for Bambara affixes and auxiliaries. The list was submitted for discussion in the "Mandelang" listserve and approved by colleagues;
- a set of rules concerning combinations of derivative affixes;
- an inventory of compounding models for Bambara;
- a Bambara-French electronic lexical database Bamadaba for the purpose of automatic morphological analysis. Charles Bailleul's dictionary (2007), taken as a basis (an electronic version of that dictionary was kindly provided by the author for the purpose of corpus building), required a serious revision, which was carried out in 2010-2011 through the efforts of Valentin Vydrin, Anna Erman, Artem Davydov, with the computational support by Kirill Maslinsky;

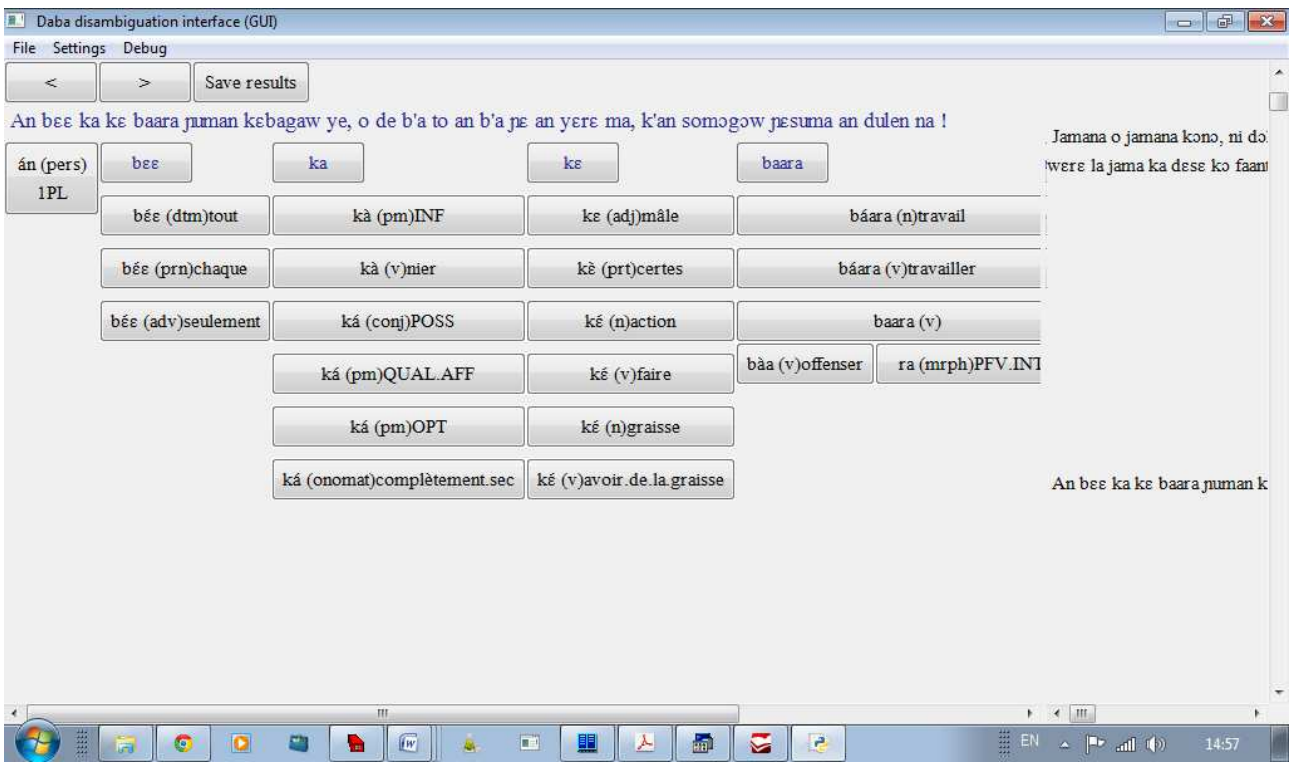

Fig. 1. A screenshot of the disambiguation interface with a Bambara non-disambiguated text.

- supplementary electronic dictionaries (for individual names, clan names, toponyms) were created on the basis of Valentin Vydrin's card index;
- electronic bibliographies (in the Toolbox format) of publications in Bambara and a bibliography of periodicals in Bambara.

In April 2012, the annotated and glossed Bambara Reference Corpus (BRC) was made accessible at the following addresses:

http://cormand.tge-adonis.fr/ (a French version)

http://maslinsky.spb.ru/bonito/run.cgi/first_form (a Russian version)

An open source version of the SketchEngine corpus search system developed by Adam Kilgariff and colleagues (NoSketchEngine) and accompanying web-concordance interface "Bonito" was adapted for the BRC.

BRC is freely accessible and provided with the following documentation (downloadable for the users):

- information about the Corpus
- a user's manual
- the ideology of the BRC
- standard glosses for Bambara affixes and auxiliary words
- POS tags used in the BRC annotation
- principles of tonal marking in the normalized Bambara texts
- a system of file naming in the BRC.

Besides, a detailed manual for the operators of the corpus-building work has been written (a French and a Russian versions); this manual is constantly updated to reflect new challenges.

2.3. The process of corpus-building

Inputting texts into the corpus represents the central part of the corpus building work. It includes the following stages.

(A) The pre-parsing stage: (a) digitalization through scanning and keyboarding (experimentation with optical recognition of Manding texts has proved to be of low effectiveness, in comparison to manual keyboarding); (b) removal of elements belonging to the descriptive metadata; (c) insertion of tags for titles, illustrations, lists, footnotes, foreign language insertions, ends of paragraphs, limits of tables, ends of lines in a verse, a list or a table; (d) insertion of the descriptive metadata. (e) For certain categories of texts containing abundant typos and errors of word segmentation, manual correction by a person with a good competence in Bambara is necessary (the original "incorrect" version of the texts is kept; it will also be made searchable, which is important for the study of orthographic norms).

(B) Bambara morphological parsing is carried out by the Daba software. The Bambara morphological parsing tools are being constantly improved in the process of text processing.

(C) The post-parsing stage consists mainly of a semi-automatic disambiguation of the text, a task that requires good proficiency in Bambara. This job is performed mainly by Daria Ogorodnikova and Elizaveta Volkova, and occasionally by Artem Davydov, Jean Jacques Méric and Valentin Vydrin. Lexemes absent in the Bamadaba are integrated into it by Valentin Vydrin or Anna Erman. Obscure fragments of the texts difficult for interpretation are sent to the Mandelang listserve for the discussion among specialists in the Bambara language.

A text in the corpus, after being subjected to all the operations, is represented at the following layers of analysis:

a) Original text corresponding exactly to the source, preserving the original orthography and misprints; b) Normalized text in Latin orthography with tonal notation. At this level, the ancient orthography is converted into the standard one, word forms are automatically identified, tone marks are inserted (if absent in the original text) or put in line with the standard model of the corpus; c) Normalized text with a complete morphemic parsing (segmented derivates and compounds); d)

Lemmatization line: each Bambara lexeme and each inflectional morpheme is provided with a French equivalent; f) Glossing line: each Bambara morpheme (of any kind) is provided with a French equivalent.

In the presentation for the users, levels (b) through (d) are combined.

The entire corpus is tone-marked, lemmatized and glossed. It includes a non-disambiguated subcorpus and a disambiguated subcorpus (the size of the latter being close to 10% of the entire BRC). A user is able either to search the entire corpus or to limit their search to the disambiguated subcorpus. There is no limit to the number of examples produced by a search (unlike in the British National Corpus), an access to the complete list of examples being often necessary for an advanced linguistic study.

2.4. Evolution of the Bambara Reference Corpus

The Members of the Bambara Corpus working group emerged in the end of 2009 had certain experience of work with the Fieldwork Toolbox software, were inspired by the success of the Russian Language National Corpus project and other similar projects, but had a rather vague idea of how one should proceed to build a million-word corpus. Two years and five months later, such a corpus appeared on the Internet.

At that moment (April 2012), it consisted of about 1,100,000 words, of these about 80,000 in the disambiguated subcorpus. A year later, the size of the BRC reached 1,498,000 words, of these about 151,000 disambiguated. It should also be mentioned that during that year, great progress was reached in the improvement of quality: numerous mistakes of analysis were corrected manually in the disambiguated subcorpus, and certain systematic errors were removed through perfection of the corpus tools (the morphological parser and the Bamadaba lexical database). Considerable progress has been reached in the corpus balancing: in April 2012, about 2/3 of the total size represented the text of the Old Testament (an electronic version offered by Charles Bailleul), in April 2013 its rate fell to 50%, the other half consisting of multifarious genres of texts: periodicals, fiction, transcriptions of oral interviews and broadcasts, vulgarization booklets, popular tales and riddles… (an exhaustive list of texts included into the BRC, together with the statistics, can be found on the site of the Corpus).

Posting of the BRC on the Internet has allowed us to begin its use in the Bambara linguistic research. During all this time, the energy of the members of the team was directed mainly to the corpus building; however, sample studies in the field of Bambara orthography usage, grammar and lexical semantics have proved that, despite its current modest size, the BRC represents a powerful tool that can potentially revolutionize the entire field of the Bambara language studies. First attempts to apply the BRC in the field of Bambara teaching at the universities (mainly at INALCO, Paris, but also at the St. Petersburg State University and École Normale Supérieure, Paris) have shown that it can also be used in classes and especially in homework, and that advanced and motivated students can take part in corpus building activities (especially disambiguating), which helps them to reach quickly a good command of the language.

It should also be mentioned that a Bambara automatic spellchecker for Open Office has been developed, on the basis of the Bamadaba lexical database, by Andrij Rovenchak. Further perfection of this spellchecker is under way.

A mid-term task of the working team is to bring the size of the BRC to 6-7 million words, with about 1 million in the disambiguated subcorpus. It is also planned to develop a parallel Bambara-French corpus of texts (in this relation, see: Rovenchak (2013)) which can be used in advanced studies of grammar and pragmatics of Bambara, and will facilitate access to Bambara texts to linguists who are not specialists in this language. One more task is the building of an oral corpus of Bambara which is indispensable for the examination of segmental, tonal and intonation-related phenomena hitherto studied mainly through elicitation. Certainly, such a corpus will be of a much smaller size (taking into account its job-intensive nature).

Another major direction is the building of corpora of other widely spoken languages of the Manding group closely related to Bambara, first of all Maninka of Guinea. A peculiarity of the Guinean Mandinka is that it is written today mainly in an original writing, "Nko", created in 1949 by a Guinean erudite and traditional scholar Solomana Kante (Oyler 2005; Vydrine 2001). Moreover, the Nko written tradition expands from Guinea to the neighboring countries where different Manding varieties are spoken; in fact, it has good perspectives to grow into a written standard for the entire Manding-speaking community of West Africa (Vydrine 2011; Vydrine 2012). First steps in this direction have been already taken: certain amount of Maninka texts is digitized, and a Nko to Latin script converter has been developed by Andrij Rovenchak.

It is also planned to set about building a text corpus of the Jula of Côte d'Ivoire, in cooperation with linguists from that country currently working on a big Jula-French dictionary.

3. Conclusions

Linguists dealing with languages of Subsaharan Africa rarely operate with text corpora of more than 100 thousand words (for exceptions, see *Introduction*). As a rule, these are individual corpora compiled by field linguists for their own purposes in the Word or Toolbox format, not available to the general public. BRC represents an attempt to overcome these limitations. Although very modest in size in comparison to reference corpora of European and some Asian languages (enjoying a long written tradition and widely used on the web), it opens a new perspective in the Manding language studies and may also present an interest from the point of view of the corpus building methodology in general.

Acknowledgements

I would like to thank the members of the Bambara Reference Corpus team for their corpus building efforts, and especially Kirill Maslinsky for his imput and feedback on this paper. I am grateful to Tatiana Nikitina who assisted me with preparing the final version of this text.

This work is related to the research strand 6 "Language resources" of the Labex EFL (financed by the ANR/CGI).

References

Asker, L. & Argaw, A. A. & Gambäck, B. & Eyassu A. S. & Habte, L. N. (2012). Classifying Amharic webnews. *Information Retrieval. 04; 12(3),* 416-435.

Aubry, N. (2010). *Changements syntaxiques dans le yoruba de la presse (1930-2010): traitement automatique d'un corpus diachronique et analyse des résultats.* Thèse de doctorat. Paris : INALCO.

Bailleul, Ch. (2007). *Dictionnaire Bambara-Français.* 3e édition corrigée. Bamako: Donniya.

Davydov, A. (2010). Towards The Manding Corpus: Texts Selection, Principles and Metatext Markup. In Guy de Pauw, Handré Groenewald & Gilles-Maurice de Schryver (eds.), *Proceedings of the Second Workshop on African Language Technology AfLaT, May 18, 2010* (pp. 59-62). Valletta, Malta [online]: http://tshwanedje.com/publications/ AfLaT2010.pdf.

De Schryver, G.-M. (2002). Web for/as corpus: A perspective for the African languages. *Nordic Journal of African Studies, 11(2),* 266-282.

Gambäck, B. (2012). Tagging and Verifying an Amharic News Corpus. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, May. ELRA. Workshop on Language Technology for Normalisation of Less-Resourced Languages.*

Hurskainen, A. (2004). Swahili language manager: A storehouse for developing multiple computational applications. *Nordic Journal of African Studies 13(3),* 363-397.

Maslinsky, K. (2012). *Daba. Pattern-based morphemic analysis toolkit.* [Software] https://github.com/maslinych/daba

Oyler, D. W. (2005). *The History of the N'ko Alphabet and its Role in Mande Transnational Identity: Words as Weapons.* Cherry Hill, NJ: Africana Homestead Legacy Press.

Rovenchak, A. (2013). Masadennin (The Little Prince in Bamana): Experimental online concordance with parallel French and English texts. *Mandenkan, 50* (in print).

Sharoff S. & Nivre J. (2011). The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge. In: *Komputernaja lingvistika i intellektual'nye tekhnologii: Po materialam Mezhdunarodnoj konferencii "Dialog" (Bekasovo, 25-29 maja 2011)* (pp. 591-604).

Sinclair, J. M. and Ball, J. (1996). *EAGLES. Preliminary Recommendations on Text Typology.* [online] http://www.fltrp.com/wyzx/07teacher/download/0706/liwz/texttyp.pdf

Vydrin, V. (2008). Glossed electronic corpora of Mande languages: A perspective that we cannot avoid. In: V.Vydrin (ed.). *Mande languages and linguistics. 2nd International Conference, St. Petersburg (Russia), September 15-17, 2008. Abstracts and Papers* (pp. 15-22). St. Petersburg: Museum of Anthropology and Ethnography.

Vydrine, V. (2001). Soulemane Kantè, un philosophe-innovateur traditionaliste maninka, vu à travers ses écrits en Nko. *Mande Studies, 3,* 99-131.

Vydrine, V. (2011). L'alternative du N'ko : une langue écrite mandingue commune, est-elle possible? In : Vold Lexander, Kristin; Lyche, Chantal, Moseng Knutsen, Anne (eds.). *Pluralité des langues pluralité des cultures : regards sur l'Afrique et au-delà* (pp. 195-204). Oslo : The Institute for Comparative Resarch in Human Culture.

Vydrin, V. (2012). Une bibliographie préliminaire des publications maninka en écriture N'ko. *Mandenkan, 48,*. 59-121.